Teilprojekt D4

Titel: Methoden zur interaktiven linguistischen Korpusanalyse von Informationsstruktur

Zusammenfassung

Im neu beantragten Projekt D4 sollen computerlinguistische Werkzeuge und Ressourcen, maschinelle Lernverfahren und statistische Methoden kombiniert werden zu einer flexiblen, interaktiven Untersuchungsmethode für Informationsstruktur (IS) in großen Korpora – sowohl für verschiedene Einzelsprachen, als auch kontrastiv auf Parallelkorpora. Der Ansatz ist charakterisiert durch maschinelles *Lernen aufgrund von interaktiver linguistischer Annotation (LAILA)* und stellt eine Ergänzung zur schon bisher im SFB entwickelten Korpus-Infrastruktur und der Annotations- und Auswertungsmethodologie dar, welche den Schwerpunkt auf relativ kleine, sorgfältig elizitierte und handannotierte Datensammlungen legt: mit der LAILA-Methode können unannotierte Korpusdaten einer raschen Exploration oder einer phänomenorientierten, kontrollierten Frequenzanalyse unterzogen werden – mit der Zielsetzung, den unumgehbaren manuellen Aufwand für die Annotation/Überprüfung von Trainings- bzw. Kontrolldaten möglichst effektiv für die linguistischen Untersuchungsziele einzusetzen.

Die linguistische IS-Forschung kann von LAILA sowohl bei der Einzelbelegsuche nach seltenen Realisationsformen in sehr großen Korpora profitieren, als auch bei der Bestimmung der Frequenzverteilung von alternativen IS-Realisierungen oder von IS-relevanten Parametern des lexikalischen, strukturellen oder Diskurs-Kontexts. Für Frequenzanalysen werden kontrollierte Stichproben erzeugt, die manuell überprüft werden und von denen ausgehend statistisch generalisiert werden kann. Frequenzdaten für große Korpora ergänzen die bisher im SFB entwickelten elizitierten Spezialkorpora zur IS komplementär: letztere kontrollieren die Verwendungskontexte für IS sorgfältig, so dass für die typologische Forschung die qualitative Vergleichbarkeit gewährleistet ist; mit Frequenzdaten kann (abhängig von dem Sprachausschnitt, den verfügbare Korpora dokumentieren) quantitativ überprüft werden, wie sich die elizitierten Realisierungsalternativen und mögliche zusätzliche Varianten in freier Sprache verteilen.

Alignierte mehrsprachige Parallelkorpus-Daten eignen sich in doppelter Hinsicht für die LAILA-Methode – einerseits als direkte Datenquelle für kontrastive Untersuchungen zur IS, andererseits zur Verbesserung der Trainingsbasis für einzelsprachliche Werkzeuge: Analyseinformation zu einer Sprache kann mit der Annotationsprojektions-Technik (Yarowsky et al. 2001) als Hilfs-Ressource für andere Sprachen ausgenutzt werden.

In der kommenden SFB-Phase stehen für D4 drei exemplarische Anwendungsszenarien im Vordergrund: (1) Für das Deutsche soll die Einzelbelegsuche und die Frequenzbestimmung der wichtigsten grammatischen Mittel zur IS-Realisierung unterstützt werden. In Kooperation mit A1 werden korpusbasiert mögliche IS-Faktoren untersucht, die die Platzierung von Relativsätzen im Deutschen (im Mittelfeld vs. extraponiert) beeinflussen. Mit Projekt C1 soll die Technik in Bezug auf die Vorfeldbesetzung durch Objekte auf dem C1-Zeitungskorpus validiert werden. (2) Auf Basis des Europarl-Korpus (Koehn 2002) mit Übersetzungen der EU-Parlamentsdebatten in 11 (bzw. 20) Sprachen soll Werkzeugunterstützung für eine kontrastive IS-Analyse bereitgestellt werden, die dann u.a. in Kooperation mit D2 für Untersuchungen zur Mikrovariation ausgenutzt werden, insbesondere zu Topikalisierungs- und Cleft-Konstruktionen. (3) Hindi dient als Beispiel für Sprachen, für die wenige Analysewerkzeuge zur Verfügung stehen. Gemeinsam mit C6 und unter Ausnutzung eines Parallelkorpus Englisch–Hindi und von Werkzeugen für das Englische soll eine Frequenzanalyse IS-relevanter Kategorien und Kontextparameter für Hindi vorgenommen werden.

3.3 Ausgangssituation des Teilprojekts

3.3.1 Stand der Forschung

Die D4-Forschungsaufgaben bauen stark auf neuere automatische Lernverfahren auf, deren Status in der linguistischen Empirie-Forschung vorab angesprochen wird.

Statistische Methoden und automatisches Lernen für linguistische Untersuchungen

Ein massiver Schwerpunkt der internationalen Sprachtechnologie-Forschung der letzten Jahre liegt auf statistischen Methoden und automatischen Lernverfahren (im folgenden ML für *Machine Learning*), mit denen sehr große Korpora technologisch ausgenutzt und analysiert werden. Linguistische Erkenntnisse spielen dabei vordergründig häufig eine untergeordnete Rolle, da bei geeigneter Wahl der Lern-Features allgemeine ML-Techniken sehr gut in der Lage sind, oberflächennahe Charakteristiken in einer großen Menge von Trainingsdaten für die Klassifikation weiterer Daten gleichen Typs auszunutzen.¹ Der Einsatz von Methoden, deren linguistischer Gehalt nicht unmittelbar klar ist, mag einer der Gründe sein, dass umgekehrt auch ML-Techniken in der theoretisch-linguistischen Forschung, also außerhalb eines vorwiegend sprachtechnologisch motivierten Kontexts, bislang relativ wenig zum Einsatz kommen² – obgleich die Techniken im Prinzip eine auf linguistische Anforderungen angepasste Analyse großer Datenmengen unterstützen könnten. Linguistische Untersuchungen, die die Frequenzverteilung von Realisierungsalternativen in Betracht ziehen (wie z.B. Bresnan et al. 2001, 2006, Bresnan/Hay 2006, Wasow et al. 2005, Bader/Häussler 2006, De Sutter et al., erscheint), könnten durch verbesserte Werkzeuge erheblich unterstützt werden, so dass ein breiteres typologisches bzw. Phänomen-Spektrum in kürzerer Zeit abgedeckt werden kann.

Es gibt zwei vordergründige Einwände gegen den „Import“ von ML-Techniken in die linguistische Empirieforschung: (1) ML-Ansätze sind auf häufige Phänomene fokussiert, da im Zuge der das Feld prägenden quantitativen Evaluationsmethodologie seltene Phänomene eine untergeordnete Rolle spielen. Für die linguistische Theoriebildung sind jedoch seltenere Phänomene von zentraler Bedeutung. (2) ML-Werkzeuge haben zwar in der Regel eine extrem große Datenabdeckung, für ihr automatisches Analyseergebnis muss jedoch immer eine gewisse Fehlermöglichkeit in Betracht gezogen werden. Linguistische Schlussfolgerungen will man jedoch auf sehr verlässliche Datenbeobachtungen stellen. Projekt D4 versucht beide Einwände durch eine geeignete Anwendung der ML-Techniken zu entkräften: Einwand (1) trifft nur zu, wenn ein gesamtes (trainiertes) Werkzeug übernommen wird. Die Trainingsverfahren können jedoch auch auf seltenere Phänomene angesetzt werden, wenn die Entwicklungsmethodologie entsprechend angepasst wird. Die Problematik der Verlässlichkeit der Analyseergebnisse (Einwand (2)) kann mit einer werkzeugunterstützten manuellen Kontrolle der abschließenden Ergebnisse vermieden werden – die natürlich einen Mehraufwand bedeutet, aber im Ergebnis immer noch eine massive Erleichterung gegenüber einem rein manuellen Verfahren bedeutet. In der Ausnutzung skalierbarer sprachtechnologischer Verfahren für die linguistische Empirieforschung steckt somit das Potential einer enormen Verbreiterung der Datenbasis ohne Aufgabe der Qualitätsstandards.

¹ In die Architektur vieler anspruchsvollerer ML-Anwendungen fließen jedoch durchaus linguistische Überlegungen ein, ebenso in die Daten-Repräsentation und die Wahl der Features. Nicht zuletzt sind ML-Ansätze auf die linguistische Annotation von Trainingsdaten angewiesen.

² Diese Beobachtung konzentriert sich auf den Einsatz sprachtechnologischer Techniken als empirisches Untersuchungsinstrument auf Korpusdaten (eine der Ausnahmen ist Philip Resniks Linguist's Search Engine <http://lse.umiacs.umd.edu>). Nicht eingeschlossen sind probabilistische Grammatikmodelle und Lernalgorithmen in der theoretischen Linguistik, die von der Sprachtechnologie-Forschung beeinflusst sind (z.B. Goldwater/Johnson 2003, Jäger 2006). Diese Ansätze haben bislang jedoch ebenfalls vergleichsweise wenig Verbreitung gefunden.

Einhergehend mit der noch vergleichsweise geringen Bedeutung von ML-Methoden für die Linguistik ist der Status von empirischen Beobachtungen zu Korpusfrequenzen für die linguistische Theoriebildung noch nicht sehr zentral. D4 vertritt die These, dass dies zum Teil auf unzureichende Verfügbarkeit bzw. Transparenz und Verlässlichkeit der entsprechenden Daten und Analysemethoden zurückzuführen ist, und will im Rahmen des SFBs den aktiven Versuch starten, mit kombinierter linguistischer und ML-Expertise die Felder anhand des besonders geeigneten Gegenstandsbereichs IS zusammenzuführen.

Werkzeuge für die Korpusanalyse von IS und IS-relevanten Kontextparametern

Viele der in den letzten 15-20 Jahren entwickelten Sprachanalysewerkzeuge können als Basis für speziellere Korpusanalysen genutzt werden. Grob können symbolische, regel-basierte Systeme vs. statistische bzw. ML-basierte Systeme unterschieden werden.³ Da IS mit allen Bereichen der Grammatik interagiert, sind Werkzeuge auf praktisch allen linguistischen Ebenen relevant. LAILA ist in der Lage, Analyseergebnisse unterschiedlicher Werkzeuge auf einfache Weise zu kombinieren und ihren IS-relevanten Gehalt zu extrahieren, so dass eine Vielzahl von Werkzeugen für den Einsatz in Frage kommen.

Regel-basierte Systeme stehen zur Verfügung u.a. für die morphologische Analyse (Karttunen et al. 1996), fürs Part-of-Speech-Tagging, für die flache syntaktische (Chunk-) Analyse (Samuelson/Voutilainen 1997, Ule/Müller 2001), für die tiefe syntaktische Analyse in verschiedenen syntaktischen Traditionen (Butt et al. 2002, Bender et al. 2002, Lin 1998) und für Koreferenz-Resolution und Identifikation von neuen vs. anaphorischen Elementen (z.B. Vieira/Poesio 1997)⁴. Daneben sind lexikalische Ressourcen zu erwähnen, die in Kombination mit verschiedenen Werkzeugen zum Einsatz gebracht werden können (WordNet, GermaNet; Hamp/Feldweg 1997; FrameNet/SALSA; Erk et al. 2003). Im Zusammenhang mit dem D4-Ansatz sollen zwei methodische Aspekte besonders hervorgehoben werden: Zum einen zeigen Eckle/Heid 1996, Evert/Kermes 2001, 2003, Kermes/Heid 2003, dass eine regelbasierte flache Syntaxanalyse erfolgreich zur interaktiven Korpusexploration eingesetzt werden kann, insbes. zur halbautomatischen Extraktion von lexikalischer Information. Die Idee hierbei ist, typische, oberflächennah identifizierbare Konfigurationen zu erfassen, die syntaktische Eigenschaften von lexikalischen Elementen charakterisieren.⁵ D4 wird die Beschränkung auf gut identifizierbare Muster als Heuristik bei der interaktiven Suche zulassen und ihren möglichen verzerrenden Einfluss auf die Datenauswahl in der Gesamtarchitektur abfedern. Hier kann auf Ergebnisse aus der komputationellen Lexikographie zur manuellen Sichtung von Zufallsstichproben zurückgegriffen werden (z.B. Evert 2004, Evert/Krenn 2005, Baroni/Evert, erscheint). Zweitens können Untersuchungen, die auf Parallelkorpus-Daten basieren, erheblich davon profitieren, wenn Parser für verschiedene Sprachen vergleichbare Repräsentationen erzeugen. Dies ist ein Kernziel größerer Kooperationsprojekte, sowohl im Rahmen der Lexikalisch-Funktionalen Grammatik (LFG; vgl. das ParGram-Projekt, Butt et al. 2002), als auch im Rahmen der Head-driven Phrase Structure Grammar (HPSG; vgl. das Grammar-Matrix-Projekt, Bender et al. 2002). Für D4 ist beabsichtigt, mit den ParGram-Grammatiken für das Deutsche (in Kooperation mit Projekt D2 (Christian Rohrer) im Stuttgarter SFB 732), das Englische, Französische und Norwegische (sowie experimentell, soweit möglich, für Hindi/Urdu; vgl. Butt/King 2002) zu arbeiten und die (parallelen) Analysen in die halbautomatische IS-Analyse einzubeziehen.

³ Im Einzelfall kommt jedoch häufig eine Kombination zum Einsatz, z.B. ein Parser mit symbolischer Grammatik und statistischer Desambiguierung.

⁴ Regel-basierte Koreferenz-Resolutionssysteme nehmen z.T. direkten Bezug auf den Informationsstatus, beispielsweise in der Tradition der Centering Theorie, vgl. Eckert/Strube 2000.

⁵ Eine verwandte Idee wird als Datenfilter im Rahmen von Arbeiten zur lexikalischen Semantik eingesetzt: Riloff/Shepherd (1997) verwenden die syntaktische Konfiguration der NP-Koordination zur Identifikation von vergleichbaren Begriffen.

Auf **statistische/ML-basierte Systeme** kann für eine Vielzahl von Analyseaufgaben zurückgegriffen werden. Die meisten bewährten Systeme arbeiten mit überwachtem Lernen, d.h. ein Korpus von Trainingsdaten wird vorab von Hand mit den gewünschten Ziel-Strukturen bzw. -Kategorien annotiert, auf dem dann das Analyse-Werkzeug trainiert wird. Zu den bekanntesten Werkzeugen dieser Typs gehören (1) Part-of-Speech-Tagger auf Basis verschiedener ML-Verfahren (Schmid 1994, Ratnaparkhi 1996, Daelemans et al. 1996, Brants 2000), (2) flache Parser oder Chunker, die zumeist Ramshaw/Marcus (1995) folgen und auf der Tagger-Technologie aufsetzen, (3) statistische Parser auf Basis eines Phrasenstruktur-Modells, wie Charniak 2000, Collins 2002 für Englisch, der Stanford-Parser (Klein/Manning 2002) für Englisch, Deutsch und Chinesisch, der Sleepy-Student-Parser (Dubey 2004) für Deutsch, (4) statistische Abhängigkeits-Parser, wie der MaltParser (Nivre et al. 2006) für Englisch, Schwedisch und Chinesisch, (5) lexikalisch-semantische Analysewerkzeuge (Erk/Pado 2006) (6) die Identifikation von neuen vs. anaphorischen Elementen und Koreferenz-Resolution (z.B. Soon et al. 2001, Ng/Cardie 2001).

Unüberwachte Lernverfahren kommen ohne hand-annotierte Trainingskorpora aus. Einerseits sind diese Verfahren von Interesse für die Grundlagenforschung zu Repräsentationsformalismen und Lernbarkeit (vgl. das DFG-geförderte Emmy Noether-Projekt PTOLEMAIOS am Lehrstuhl Kuhn); für bestimmte Aufgaben lassen sich aber durchaus sehr gute Analyseergebnisse erzielen, z.B. für Part-of-Speech-Tagging (Brill 1993), Wortbedeutungs-Desambiguierung (Schütze 1998) und für Wort-Alignierung auf Parallelkorpora bei einer gegebenen Satz-Alignierung (Brown et al. 1993). Letzteres Verfahren wird in D4 als Vorbereitungsschritt auf die verwendeten Parallelkorpora angewendet.

Ein wichtiger Aspekt für D4 ist die **Kombination von bestehenden Werkzeugen** für eine komplexere oder speziellere Aufgabe oder zur Verbesserung der Qualität (vgl. z.B. Henderson/Brill 1999). Hier wurden mit ML-Techniken Fortschritte erzielt, die sich auf die D4-Problematik übertragen lassen. Als Beispiel sei die Aufgabe der semantischen Rollenzuweisung erwähnt, für die Pradham et al. (2005) erhebliche Verbesserung dadurch erzielen, dass sie der Aufgabe drei verschiedene Parser (einen flachen, einen tiefen statistischen und einen regelbasierten) zugrundelegen und die Ergebnisrepräsentationen kombinieren.

Schwach überwachte Lernverfahren

Die Idee dieses Mittelwegs besteht darin, den manuellen Annotierungsaufwand für die Erstellung von Trainingsdaten zu verringern und stattdessen implizite Information in unannotierten Daten auszunutzen (vgl. auch „klassische“ halbautomatische Korpusannotation z.B. Brants 1999). Im **Active-Learning**-Ansatz (s.z.B. Thompson et al. 1999) werden die Daten, die dem Annotierer als nächstes präsentiert werden, und auf deren Basis ein zyklisch verbessertes Training stattfindet, systematisch ausgesucht. In *Uncertainty-Sampling*-Ansätzen (Cohn et al. 1995) werden diejenigen Daten aus dem Pool von unannotierten Sätzen ausgesucht, für die die Unsicherheit des aktuellen Werkzeugs am größten ist – sprich, für die die Wahrscheinlichkeitsverteilung über die verschiedenen Analysehypothesen am flachsten verläuft.⁶ Für *Uncertainty Sampling* wird nur *ein* Modell (das sukzessive immer neu trainiert wird) benötigt; das Verfahren lässt sich jedoch auch mit einem *Ensemble* von verschiedenen Modellen verwenden, die eine unterschiedliche Architektur oder lediglich unterschiedliche Features haben können. Die Ungewissheit wird dann für das kombinierte Analyseergebnis berechnet.⁷ Ensemble-basiertes aktives Lernen führt generell zu besseren Ergebnissen als Einzelmodell-Lernen, wobei mit der *Bagging*-Methode ein Ensemble auch aus einem

⁶ Eine sehr effektive Vereinfachung dieses Kriteriums wird von Baldrige/Osborne (2003) im Rahmen einer Studie zur HPSG-Desambiguierung vorgeschlagen: mit der *Lowest Best Probability*-Methode setzen sie die Konfidenz eines Modells für eine Analyse mit der Wahrscheinlichkeit der besten Alternative gleich.

⁷ Eine alternative Ensemble-Methode arbeitet mit *Query-by-Committee* (Seung et al. 1992), bei dem die Daten aufgrund von unterschiedlichen Analysevorhersagen (also unterschiedlichen „Gewinnern“) ausgesucht werden.

Einzelmodell erzeugt werden kann. Dabei werden für verschiedene Modellinstanzen unterschiedliche Zufalls-Stichproben (mit Zurücklegen) aus den Trainingsdaten genommen. Mit dieser Technik „sehen“ die verschiedenen Modellinstanzen jeweils andere Teile der Trainingsdaten, verhalten sich also bei niedrig-frequenten Ereignissen unterschiedlich. Becker/Osborne (2004) zeigen anhand von Active-Learning-Experimenten zum Training von Parsern, dass die *Bagging*-Methode (als Teil eines Zwei-Stufen-Modells) vor allem für Ereignisse mit niedriger Auftretenshäufigkeit einen Vorteil bringt.

Ein Gutteil der Forschung zum Active Learning arbeitet lediglich mit einer simulierten sukzessiven Annotation, indem ein bestehendes, komplett annotiertes Korpus auf unterschiedliche Art erschlossen wird. Eine reale Anwendung von Active-Learning-inspirierten Techniken, wie in D4 geplant, kann somit zu interessanten Erkenntnissen führen.

Bootstrapping- und **Co-Training**-Ansätze gehen nur von einem einmalig manuell annotierten „Saat“-Trainingskorpus aus und wählen im Anschluss aus einem unannotierten Pool Trainingsdaten aus, für die mit dem aktuellen Modell auch die Zielanalyse automatisch bestimmt wird. Die Datenauswahl basiert auf vergleichbaren Techniken wie beim Active Learning – hier allerdings auf Basis von *größtmöglicher* Konfidenz in die automatisch zugewiesenen Analyse. In der Sprachtechnologie hat sich diese Technik bei „flacheren“ Aufgaben als sehr effektiv herausgestellt (Yarowsky 1995, Blum/Mitchell 1998; vgl. zum Parsing Steedman et al. 2003).

Arbeiten auf Parallelkorpora: die Annotationsprojektions-Technik

Eine ML-Technik, die gewissermaßen als schwach überwacht bezeichnet werden kann, nutzt Parallelkorpora (auf denen unüberwacht eine statistische Wort-Alignierung bestimmt wurde), um Analyse-Werkzeuge für eine Sprache zu trainieren, für die keine annotierten Trainingsdaten vorliegen. Dazu wird die englische Seite des Korpus mit einem existierenden Analysewerkzeug (z.B. einem Chunker) analysiert, und die Wortentsprechungen zur anderen Sprache werden genutzt, um die Analysen auf die andere Sprache zu „projizieren“ (Yarowsky et al. 2001). Diese Technik wurde bereits für Part-of-Speech-Tagging, Chunking, morphologische Analyse, Dependenz-Parsing (Hwa et al. 2002) und für Aspekte der lexikalischen Semantik eingesetzt (s.z.B. Pado/Lapata 2005 zu einem syntaktisch informierten Projektionsmodell) und ist für vereinfachte Analyseaufgaben (z.B. reduzierte Tagsets) erstaunlich effektiv. Im Rahmen von D4 bietet sich die Annotationsprojektions-Technik sehr als ergänzende Quelle für Lern-Features an, insbesondere wenn mehrsprachige Parallelkorpus-Daten zur Verfügung stehen und so manuelle Annotationsarbeit gleich für mehrere andere Sprachen ausgenutzt werden kann.

3.3.2 Eigene Vorarbeiten

Das D4-Projekt führt den thematischen Schwerpunkt Informationsstruktur (IS) aus früheren Arbeiten von **Jonas Kuhn** und beide von ihm in den letzten Jahren vertieften korpus-orientierten Ansätze – linguistisches *Knowledge Engineering* für Grammatiken mit breiter Abdeckung einerseits und die Entwicklung von statistischen/ML-Methoden andererseits – zusammen.

Im Zentrum von Kuhns früheren Arbeiten zur IS standen (computer-)linguistische Fragen zu Struktur und Bedeutung von IS, vorwiegend in Bezug auf syntaktische und prosodische Aspekte der IS-Realisierung im Deutschen (Kuhn 1996a,b, 1999a, 2001c, Dogil et al. 1997). Die Frage der Korpusverteilung der untersuchten Phänomene wurde in diesen Arbeiten nicht direkt angesprochen, bildeten jedoch ebenso wie das Desiderat einer eingehenderen empirischen Klärung der Verwendungsbedingungen für IS-Realisierungen eine wichtige Motivation dafür, dass sich die Arbeiten Kuhns in der Folge stark auf linguistisch fundierte, dabei jedoch auf breite Abdeckung zielende Ansätze in der komputationellen Syntax ausrichteten. Ein langfristiges Ziel

war immer, mit verbesserten computerlinguistischen Werkzeugen auf das Thema IS zurückkommen zu können.⁸

Im Rahmen des multilingualen ParGram-Projekts (vgl. u.a. King et al. 2004, 2005) zielen eine Reihe von Kuhns Beiträgen auf die Verwendung der sorgfältig entwickelten Grammatiken zur Verbesserung der Ausgangssituation für andere (computer-)linguistische Aufgaben, z.B. die Erstellung von Lexika aus Korpora (Kuhn et al. 1998), die syntaktische Annotation von Korpora (Zinsmeister et al. 2002) oder Anaphernresolution (Asher et al. 2004, Denis/Kuhn 2006). D4 wird die dabei angewandte korpus-gerichtete Methodologie in Hinblick auf die linguistische Verwertbarkeit und die statistische Interpretierbarkeit erheblich verfeinern.

Ein weiterer Schwerpunkt in Kuhns Arbeiten ist der Einsatz statistischer Methoden und ML-Verfahren im Rahmen von linguistisch fundierter Sprachverarbeitung (z.B. Riezler et al. 2000, Kuhn 2002a, 2004, Forst et al. 2005), darunter gezielte Experimente zur raschen Bereitstellung von Werkzeugen für die Korpusexploration in wenig erschlossenen Sprachen, wie der Maja-Sprache Q'anjob'al (Kuhn/Mateo-Toledo 2004). Im Rahmen des PTOLEMAIOS-Projekts werden zum einen die Grammatikinduktion und -bootstrapping aus der impliziten Korrespondenz-Struktur in Parallelkorpora untersucht (Kuhn 2004a, Kuhn/Jellinghaus 2006, Hopkins/Kuhn 2006a,b), zum anderen Syntax-basierte Verfahren der statistischen maschinellen Übersetzung (laufende Arbeiten von Hopkins/Kuhn). Auf das in diesem Rahmen entwickelte und implementierte generative Modell der *Hierarchical Labeling Processes (HLPs)* wird in Abschnitt 3.4.2 zu den Methoden detaillierter eingegangen (Fußnote).

Mit der Kombination von linguistischer Expertise zu IS und computerlinguistischem Know-How zu den relevanten sprachtechnologischen Techniken befindet sich D4 in einer sehr guten Ausgangsposition für den Versuch, eine spezielle Untersuchungsmethodologie zu entwickeln und in den SFB einzuführen.

3.3.3 Liste der publizierten einschlägigen Vorarbeiten

Begutachtete Beiträge in Zeitschriften und Sammelbänden

- King, Tracy H., Stefanie Dipper, Anette Frank, Jonas Kuhn und John Maxwell (2004). Ambiguity management in grammar writing. *Research on Language and Computation*, 2 (2): 259-280, Dordrecht: Kluwer Academic Publishers.
- King, Tracy, H. Martin Forst, Jonas Kuhn und Miriam Butt (2005). The Feature Space in Parallel Grammar Writing. Special issue on Shared Representation in Multilingual Grammar Engineering, 3 (2): 139 - 163, Dordrecht: Kluwer Academic Publishers.
- Kuhn, Jonas (1999a). The syntax and semantics of split NPs in LFG. In F. Corblin, C. Dobrovie-Sorin und J.-M. Marandin (Eds.), *Empirical Issues in Formal Syntax and Semantics 2, Selected Papers from the Colloque de Syntaxe et Sémantique à Paris (CSSP 1997)*, The Hague: Thesus, 145–166.
- Kuhn, Jonas (2001c). Resource sensitivity in the syntax-semantics interface and the German split NP construction. In T. Kiss und D. Meurers (Eds.), *Constraint-Based Approaches to Germanic Syntax*. Stanford: CSLI Publications, 177–215.

Kongressbeiträge (mit Begutachtung der vollständigen Beiträge)

- Hopkins, Mark und Jonas Kuhn (2006a): Exploring the Potential of Intractable Parsers. In: *Proceedings of ACL/COLING 2006*.
- Kuhn, Jonas (1996a). An underspecified HPSG representation for information structure. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, 670–675.
- Kuhn, Jonas (2004a). Experiments in Parallel-Text Based Grammar Induction. In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, Barcelona.
- Kuhn, Jonas, Judith Eckle und Christian Rohrer (1998). Lexicon acquisition with and for symbolic NLP-systems – a bootstrapping approach. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC98)*, Granada, Spain, 89–95.

⁸ Vgl. z.B. die Diskussion des bidirektionalen optimalitätstheoretischen Grammatikmodells in Kuhn 2003c, Kap. 5 und die Ableitung von Präferenzen zwischen Stellungsvarianten im Deutschen aus Korpusdaten in Kuhn 2002a.

- Kuhn, Jonas und B'alam Mateo-Toledo (2004). Applying Computational Linguistic Techniques in a Documentary Project for Q'anjob'al (Mayan, Guatemala). In Proc. of LREC-2004, Lissabon.
- Kuhn, Jonas und Michael Jellinghaus (2006): Multilingual parallel treebanking: a lean and flexible approach. In: Proc. of LREC-2006. Genoa, Italy.
- Riezler, Stefan, Detlef Prescher, Jonas Kuhn, and Mark Johnson (2000). Lexicalized Stochastic Modeling of Constraint-Based Grammars using Log-Linear Measures and EM Training. Proc. of ACL-2000, Hongkong, 480-487.

Monographie

- Kuhn, Jonas. 2003c. Optimality-Theoretic Syntax—A Declarative Approach. Stanford, CA: CSLI Publications.

Beiträge zu Konferenzen und Workshops

- Asher, N., P. Denis, J. Kuhn, E. Larson, E. McCready, A. Palmer, B. Reese, L. Wang (2004). Extracting and Using Discourse Structure to Resolve Anaphoric Dependencies: Combining Logico-Semantic and Statistical Approaches. In TALN Proceedings (SDRT Workshop), Fès, Marokko, 515–524.
- Denis, Pascal und Jonas Kuhn (2006). Applying an LFG Parser in Coreference Resolution -- Experiments and Analysis. In Proceedings of the LFG 2006 Conference, Konstanz.
- Dogil, Grzegorz, Jonas Kuhn, Jörg Mayer, Gregor Moehler und Stefan Rapp (1997). Prosody and discourse structure: issues and experiments. In Proceedings of the ESCA Workshop on Intonation: Theory, Models and Applications. Athen, 99–102.
- Forst, Martin, Jonas Kuhn and Christian Rohrer (2005). Corpus-based learning of OT constraint rankings for large-scale LFG grammars. In Proceedings of the LFG 2005 Conference, Bergen, Norway.
- Hopkins, Mark und Jonas Kuhn (2006b): A Framework for Incorporating Alignment Information in Parsing. In: Proceedings of the EACL 2006 Workshop on Cross-Language Knowledge Induction.
- Kuhn, Jonas (1996b). Context effects on interpretation and intonation. In D. Gibbon (Ed.), Natural Language Processing and Speech Technology. Results of the 3rd KONVENS Conference (Konferenz "Verarbeitung Natürlicher Sprache"), Berlin: de Gruyter, 186–198.
- Kuhn, Jonas (2002a). Corpus-based Learning in Stochastic OT-LFG – Experiments with a Bidirectional Bootstrapping Approach. In Proceedings of the LFG 2002 Conference, Athen, 239–257.
- Zinsmeister, Heike, Jonas Kuhn und Stefanie Dipper (2002). Utilizing LFG Parses for Treebank Annotation. In Proceedings of the LFG 2002 Conference, Athen.

3.4 Planung des Teilprojekts (Ziele, Methoden, Arbeitsprogramm)

3.4.1 Projektziele und Methodenübersicht

Der Kern von D4 ist die Entwicklung, Implementierung und exemplarische Umsetzung einer Methode zur Analyse von IS in Korpora – mit der Idee des *Lernens aufgrund von interaktiver linguistischer Annotation (LAILA)*. Wir können **zwei Arten von Projektzielen** unterscheiden: es gilt zum einen, die Analysemethode **technisch-formal** auszuarbeiten, für die IS-Forschung umzusetzen und zu implementieren und im weiteren Verlauf so zu verfeinern, dass der benötigte manuelle Annotations- bzw. Kontroll-Aufwand möglichst effektiv eingesetzt wird. Details hierzu werden im folgenden weiter ausgeführt. Zum anderen wird **interdisziplinär-methodologisch** angestrebt, Verfahren aus der Sprachtechnologie auf breiterer Front für die linguistische IS-Forschung zugänglich zu machen und exemplarisch zu zeigen, dass damit ein anderer Zugang auf größere Sammlungen von Korpusdaten eröffnet wird, der möglicherweise bestehende empirische Methoden komplementiert.⁹

Zu den Forschungsaufgaben für die Umsetzung der LAILA-Methode zählt erstens die Entwicklung bzw. Anpassung von Werkzeugen für die IS-Analyse: so werden auf Basis von

⁹ Diese zweite Gruppe von Zielen hängt einerseits von einer erfolgreichen Umsetzung der erstgenannten Ziele ab, der Erfolg ist andererseits jedoch wesentlich offener und weniger steuerbar. Im Minimalfall dürften die entwickelten Methoden in jedem Fall für interessierte Anwender wertvoll sein; im günstigen Fall könnte ein leicht zu verstehendes und zu bedienendes Analysewerkzeug mit einem größeren Verbreitungsgrad entstehen.

existierenden flachen und tiefen syntaktischen Analyse-Werkzeugen automatische Klassifikatoren für strukturelle Aspekte der IS-Realisierung entwickelt, z.B. für Linearisierungsalternativen und Konstruktionen wie Topikalisierung und Cleft. Zur Bestimmung des Informationsstatus (gegeben vs. neu) wird eine Abwandlung bestehender Koreferenz-Resolutions-Algorithmen eingesetzt. Die zweite, zentralere Aufgabe besteht in der Entwicklung einer interaktiven Methodologie zum Werkzeug-Training und zur *halb*automatischen Korpusuntersuchung unter Einbeziehung bestehender hand-annotierter Ressourcen. Dazu sind Techniken des schwach überwachten automatischen Lernens speziell auf die hohen Qualitätsanforderungen der linguistischen Analyse anzupassen, wobei ausgenutzt werden kann, dass die Systembenutzer – Linguisten bei der Datensichtung – eine hohe Bereitschaft und Kompetenz zu manuellen Beiträgen mitbringen.

Das Ziel der interaktiven Methode ist es, den unumgehbaren **Aufwand der manuellen Annotierung bzw. Nachkorrektur von Korpusbelegen** möglichst **effektiv für die linguistische Forschung** zu gestalten – in Bezug auf die jeweilige Fragestellung an das Korpus. So lohnt sich für eine explorative Korpusuntersuchung zu einem eher seltenen Phänomen (z.B. Spalt-Konstruktionen à la „*Auch Althasen habe ich noch keine gesehen*“¹⁰) keine aufwändige Handannotierung der relevanten Faktoren; eine heuristische Suche nach zu erwartenden Oberflächen-Korrelaten mit anschließender manueller Ausfilterung erscheint sinnvoller. Unterstützung durch automatische Analysewerkzeuge, beispielsweise einen Parser, kann sich jedoch als sehr effektiv erweisen. **Für eine Korpusuntersuchung zu Frequenzverteilungen** – beispielsweise zum Vergleich von Objekt-Topikalisierungen bei belebtem vs. unbelebtem Subjekt – darf nicht die Recall-Maximierung bei der Phänomensuche im Mittelpunkt stehen, sondern das abschließend erzielte quantitative Ergebnis muss **repräsentativ für die Verteilung im Gesamtkorpus** sein.¹¹ Dies erfordert einen höheren manuellen Aufwand, der jedoch wiederum an unterschiedlichen Stellen eingesetzt werden kann. Die geplante Architektur wird einen erheblichen Teil dieses Aufwands in die Konfidenz-Bewertung von wiederverwertbaren Einzel-Analysewerkzeugen lenken und daneben statistische Methoden zur Vorhersage der Konfidenz bei kombinierten Korpus-Anfragen anbieten.

Die drei größeren **exemplarischen Anwendungsszenarien** (zum Deutschen, zu Mikrovariation in europäischen Sprachen und zu Hindi) – jeweils im Austausch mit anderen Teilprojekten – wurden bereits in der Zusammenfassung vorgestellt. Offenkundige Vorteile der Kooperationen liegen nicht nur im Testen und Bekanntmachen der Methode, sondern auch im Zugriff auf das Know-How und auf bestehende Ressourcen der Kooperationspartner – so bietet beispielsweise das bestehende Korpus von C1 zu Objekt-Topikalisierungen eine hervorragende Möglichkeit, die LAILA-Methode zu validieren.

Über die drei Hauptszenarien hinaus streben wir eine leichte Erweiterbarkeit der Werkzeuge auf andere Korpora/Sprachen an. Die D4-Kernmethodologie soll im Projektverlauf gerade auch auf Sprachkorpora erweitert werden (evtl. unter Rückgriff auf verfügbare Ressourcen wie Audiobücher), für die das Sprachsignal und/oder eine prosodische Annotation vorliegen. Hierzu sollen auch über den SFB hinaus bestehende Kontakte vertieft werden, u.a. mit Projekt A1 (Grzegorz Dogil/Hans Kamp) im Stuttgarter SFB 732 und mit der korpus-basiert arbeitenden Dolmetsch-Wissenschaftlerin Barbara Ahrens an der FH Köln (vgl. Ahrens 2003).

¹⁰ http://www.niedersaechsischer-jaeger.de/Forum/board_entry.php?id=267; gefunden durch Internet-Suche nach der Wortsequenz „noch keine gesehen“.

¹¹ Dieses Desiderat ist unabhängig von der Frage, ob das Gesamtkorpus repräsentativ für die Sprache/eine bestimmte Sub-Sprache ist (was für effektiv zur Verfügung stehende Korpora nicht der Fall sein mag – mit der Methode lassen sich jedoch Frequenzverteilungen in unterschiedlichen Korpora oder Subkorpora relativ leicht vergleichen, so dass man abschätzen kann, welche Ergebnisse u.U. trotzdem aussagekräftig sind).

3.4.2 Methoden und Architektur

Analysewerkzeuge

Der grundsätzliche Aufbau der LAILA-Architektur ist wie folgt: wir unterscheiden Hintergrund- und Kern-Analysewerkzeuge (Hintergrund- und Kern-AWs). Kern-AWs sind in jedem Fall statistisch/ML-basiert und werden im Rahmen von LAILA auf Korpusdaten trainiert,¹² während es sich bei den Hintergrund-AWs um existierende AWs (z.B. Part-of-Speech-Tagger und Parser) handelt, die lediglich zur Unterstützung des Kern-Trainings dienen, also zur Erzeugung von Lern-Features. Die trainierten Kern-AWs nehmen diskriminative Klassifikationsaufgaben wahr, an denen wir für korpus-basierte IS-Studien interessiert sind, d.h. sie identifizieren IS-relevante lexikalische, strukturelle und Diskurskontext-Parameter (z.B. syntaktische Konstruktionen wie Topikalisierung oder den Informationsstatus von NPs).¹³ Nach einer Phase des Experimentierens sollen für besonders zentrale Analyseaufgaben in der IS-Forschung vordefinierte und -trainierte Kern-AWs bereitgestellt werden, die im Rahmen der interaktiven Korpusuntersuchung durch studienspezifische Kern-AWs ergänzt werden können.¹⁴

Korpusdaten, Training und Kontrolle

Im allgemeinen Fall wird der LAILA-Ansatz auf eine Kombination von Korpusdaten unterschiedlichen Typs angewendet:

1. Existierende hand-annotierte Daten der Sprache L_1 mit Annotation der relevanten IS-Unterscheidungen (z.B. vollständig annotierte QUIS-Daten),
2. Weitere hand-annotierte Korpusdaten der Sprache L_1 mit verlässlichen grammatischen Annotationen anderer Art (z.B. eine Baumbank wie die deutsche TIGER-Baumbank),
3. Unannotierte Korpusdaten der Sprache L_1 ,
4. Unannotierte Parallelkorpusdaten (mit Satzalignierung) der Sprache L_1 mit Übersetzungsentsprechungen in den Sprachen $L_2, L_3, \dots, L_n$ (z.B. das Europarl-Korpus).

Auf alle Daten in Sprache L_1 werden die Hintergrund-AWs (automatische Part-of-Speech-Tagger, diverse Parser etc.) für L_1 angewandt. Auf die Parallelkorpusdaten werden zusätzlich entsprechende AWs für die anderen Sprachen angewandt, außerdem wird eine statistische Wort-Alignierung für alle Sprachpaare $L_1-L_2, L_1-L_3, \dots, L_1-L_n$ bestimmt. Aus den Hintergrund-

¹² Wird ein Kern-AW gewünscht, das systematisch eine Teilanalyse eines bestehenden regel-basierten Systems kopiert (beispielsweise die Identifikation einer Topikalisierungsstruktur, für die die bestehenden Parser zuverlässige Ergebnisse liefern) kann dies über eine sehr kleine Menge von Trainingsdaten vermittelt werden. Der Vorteil besteht darin, dass auf Basis der Active-Learning-Technik auch Beispiele abgefragt werden, die einer der Parser nicht abdeckt, so dass das resultierende AW robuster wird.

¹³ Kern-AWs können selbst wieder zur Erzeugung von Lern-Features für das Training von anderen Kern-AWs eingesetzt werden. So kann z.B. für die Klassifikation von Kontrastierungs-Kontexten auf die einfachere Aufgabe der Informationsstatus-Klassifikation zurückgegriffen werden. Zur Berücksichtigung von hierarchischen Abhängigkeiten zwischen AWs können wir direkt auf das im PTOLEMAIOS-Projekt entwickelte und in Java implementierte Modell der *Hierarchical Labeling Processes (HLPs)* zurückgreifen (Hopkins/Kuhn 2006a). Dieses verallgemeinerte Bottom-up-Parsing-Modell ist ein generatives statistisches Modell, das es erlaubt, beliebige Klammerungs- und Labeling-Prozesse über einen String zu beschreiben und für jede Einzelentscheidung einen ML-Klassifikator einzusetzen (wobei die Wahl des ML-Verfahrens offen ist). Zusätzlich können harte Constraints über die hierarchischen Strukturen (z.B. den Baumcharakter) und die Beziehung verschiedener Strukturen untereinander formuliert werden. Eine große Stärke des HLP-Ansatzes ist, dass es praktisch keine Beschränkung für die verwendbaren Lern-Features gibt, da die Suche nicht mit dynamischer Programmierung realisiert ist und so keine Kapselung von Teilstrukturen erzwungen wird, über die nicht hinweggegriffen werden darf.

¹⁴ Die Spezifikation von Klassifikationsaufgaben für die vordefinierten AWs wird in enger Abstimmung mit Projekt D1 geschehen (das im Rahmen der „klassischen“ Korpusannotation auf eine halbautomatische Annotations-Unterstützung zuarbeitet) und soll dem Prinzip folgen, dass jede Einzelaufgabe möglichst einfach gehalten wird – idealerweise als binäre Entscheidung oder als Entscheidung zwischen sehr wenigen Kategorien. Dies erleichtert das Experimentieren mit unterschiedlichen ML-Ansätzen und maximiert den Wiederverwendungswert der AWs, so dass die Erstellung von Trainingsdaten und die manuelle Konfidenz-Analyse nicht wiederholt werden müssen. Für die komplexere Suche werden die einzelnen Kern-AWs kombiniert.

Analysen werden schematisch ein große Anzahl von Lern-Features erzeugt, die in den ML-Verfahren zu gewichten sind.

Zur Verdeutlichung des Trainingsprozesses soll hier zunächst ein voll überwachter Ansatz skizziert werden. In diesem Fall könnte im Prinzip ausschließlich mit Daten vom Typ 3 gearbeitet werden. Ein Teil der unannotierten Daten wird zufällig ausgewählt und in einen Trainings- und einen Kontrollbereich unterteilt.¹⁵ Der Rest der unannotierten Daten ist unser Untersuchungskorpus. Nehmen wir an, wir sind an der Verteilung von belebten vs. unbelebten Objekt-NPs im Vorfeld interessiert. Dazu muss zunächst ein Kern-AW (a) trainiert werden, das NPs identifiziert, des weiteren Kern-AWs (diskriminative Klassifikatoren), die für eine NP entscheiden, ob sie (b) im Vorfeld steht, (c) ein Objekt ist und (d) belebt vs. unbelebt ist. Für das Training von (a) werden beispielsweise in ca. 400 Sätzen alle NPs annotiert (in diesem Fall am einfachsten durch Korrektur der Analysen eines automatischen Parsers oder Chunkers); zur Kontrolle der Zuverlässigkeit wird vorab auch ein Kontrollkorpus von ca. 100 Sätzen annotiert. Nach dem Training kann am Kontrollkorpus Precision und Recall bestimmt werden. Für (b)-(d) werden in einem anderen Teil des Trainingskorpus einige der mit (a) automatisch klassifizierten NPs entsprechend der jeweiligen Aufgabe detaillierter annotiert. Im Kontrollkorpus werden die bereits annotierten NPs verfeinert annotiert; wenn die Anzahl der Instanzen für bestimmte Ereignistypen im Kontrollkorpus zu klein wird, müssen weitere Daten aus dem Kontrollbereich nachannotiert werden. Die diskriminativen Klassifikatoren werden nun auf den jeweiligen Daten trainiert. Anhand des Kontrollkorpus lässt sich für jedes einzelne Kern-AW die Zuverlässigkeit bestimmen. Auf dieser Basis werden die AWs auf das bisher ungesehene Untersuchungskorpus angesetzt, (1) als Kombination (a)/(b)/(c)/(d) (also für Objekt-NPs im Vorfeld) und außerdem (2) als (a)/(b)/(d) (also für Objekt-NPs) und (3) als (a)/(c)/(d) (für Vorfeld-NPs). Es ergeben sich jeweils zwei Belegmengen, für die Belebtheit und Unbelebtheit vorhergesagt wird und die für die Frequenzabschätzung eingesetzt werden können. Da aus den Kontrollbeurteilungen für die Einzel-AWs Fehlerabschätzungen vorliegen, kann für die kombinierte Anwendung eine Fehlererwartung hochgerechnet werden. Für Stichproben aus den Belegmengen kann andererseits manuell überprüft werden, wie groß die tatsächliche Fehlerquote ist (echte vs. falsche positive Daten und echte vs. falsche negative Daten), so dass sich vorhersagen lässt, mit welcher Wahrscheinlichkeit die beobachtete Frequenzverteilung ein reiner Zufallseffekt war.

Für einen interaktiven, schwach überwachten Trainingsansatz, in dem der Annotierungsaufwand reduziert werden soll, gibt es eine Reihe von Ansatzpunkten: **(i)** Die Kontrolldaten werden bevorzugt mit Daten vom Typ 1 und 2 bestückt, also mit bereits annotierten Daten; wenn sehr viele annotierte Daten vorliegen, können diese selbstverständlich auch als Trainingsdaten genutzt werden. **(ii)** Die zu annotierenden Trainingsdaten werden mit oberflächennahen Suchausdrücken vorgefiltert (für die Kontrolldaten darf dies nicht geschehen, damit keine Verzerrung entsteht). **(iii)** Mit Active-Learning-Methoden werden die informativsten Trainingsdaten ermittelt und gezielt annotiert; durch den Rückgriff auf unterschiedliche Hintergrund-AWs und/oder unterschiedliche Sprachpaare bei der Verwendung von Daten vom Typ 4 ergibt sich eine gute Basis für Ensemble-Methoden. Laufende Kontrollüberprüfungen minimieren den verzerrenden Effekt der Active-Learning-Datenauswahl.¹⁶ **(iv)** Ideen aus dem Bootstrapping und dem Active Learning können kombiniert werden: solange sich klare Verbesserungen in der Kontroll-Evaluation erzielen lassen, werden die Daten mit der jeweils geringsten Modell-Ungewissheit mit der automatischen Analyse unbesehen in die Trainingsdaten übernommen; wenn sich kein Fortschritt mehr erzielen lässt, werden einige Active-Learning-Schritte zwischengeschaltet usw. **(v)** Typische Trainingsdaten für eine Klassifikation lassen sich z.T. sehr leicht finden, indem man sich systematisch auf eine

¹⁵ Selbstverständlich kann an geeigneten Stellen auch die Kreuz-Validierungstechnik eingesetzt werden.

¹⁶ Für Kern-AWs, deren Aufgabe nahe an existierende Hintergrund-AWs angelehnt ist (wie im obigen Beispiel das NP-AW) müssen nur wenige Beispiele annotiert werden, um den gewünschten Effekt zu erreichen; in realen Szenarien fällt vermutlich eine Vielzahl der benötigten Kern-AWs in diese Kategorie.

Teilmenge von zweifelsfrei positiven und negativen Instanzen konzentriert, die sich oberflächennah identifizieren lässt (Beispiel: für kasusabhängige Entscheidungen werden ambige Kasusformen ignoriert; vgl. Evert 2004).¹⁷

Den Daten aus (mehrsprachigen) Parallelkorpora kommt wie bereits ausgeführt eine Doppelrolle zu. Ihre technisch motivierte Verwendung zur Projektion von Annotationen über Sprachen hinweg kann sehr gut in die interaktive ML-Architektur integriert werden; das Maß der sprachübergreifenden Verwertbarkeit muss experimentell ermittelt werden. Sollten sich Parallelkorpus-Daten für die LAILA-Methodologie bewähren, bietet es sich an, manuelle Arbeitsschritte (wie im Active Learning) schwerpunktmäßig auf solchen Daten vorzunehmen, da sich über die Projektion die Verwertbarkeit multipliziert; denkbar ist auch eine Parallel-Annotation von Übersetzungsdaten für mehrere Sprachen gleichzeitig – unterstützt durch eine mehrsprachig sensitive Active-Learning-Methodologie. Der Annotationsaufwand für ein mehrsprachiges Übersetzungstupel fällt deutlich geringer aus als die Summe entsprechender Einzelannotationen, da das Hineindenken in den Satzinhalt und -kontext nur einmal anfällt (vgl. Kuhn/Jellinghaus 2006).

Eine der Herausforderungen für D4 ist es, die für das Training und die Kontroll-Beurteilung der unterschiedlichen Kern-AWs benötigten Korpusdaten und deren Annotationsstatus (unannotiert / automatisch annotiert mit AW X, Variante N / hand-annotiert von Person Y im Rahmen von Prozess Z) zu verwalten und dabei Buch zu führen über die Konfidenz der AWs und mögliche Verzerrungen gegenüber der intendierten Klassifikationsaufgabe.¹⁸

3.4.3 Arbeitsprogramm

Die Aufgaben für D4 lassen sich folgendermaßen untergliedern:

A. Datenaufbereitung und Anwendung von Hintergrund-Analysewerkzeuge

2007/08: Formatspezifikation und -konversion; Vorprozessierung der Korpora (Zeichen-Kodierungsfragen, Tokenisierung usw.); Auswahl und Anwendung der Hintergrund-Analysewerkzeuge; Infrastruktur für Annotationsprojektions-Technik; 2009/10: Automatisierung der Hintergrund-Analyse zur verbesserten Erweiterbarkeit der Datenbasis.

B. Kern-Analysewerkzeuge: Spezifikation, Maschinelle Lerntechniken

2007/08: Trainingsexperimente von IS-relevanten Klassifikatoren mit unterschiedlichen ML-Techniken; Feature-Selektion; Konzeptuelle Einbindung in das Hierarchical Labeling Process-Modell und Implementierung der benötigten Schnittstellen; 2009: Spezifikation von Standardtypen von Klassifikatoren; 2010/11: Optimierung des Trainings für spezielle AWs und Kombinationen von AWs; systematische Vergleichsexperimente.

C. Schwach überwachte Lernverfahren, statistische Fehleranalyse und Generalisierung

2007-09: Experimente mit oberflächennahen Filterausdrücken und unterschiedlichen Varianten des Active Learnings/Bootstrapping; theoretische Arbeiten zu Konfidenz-/Unsicherheitsbewertung und statistischer Vorhersage der erwarteten Fehlerquote; 2010/11: Verfeinerung und Optimierung der bewährtesten Techniken; Erweiterung auf eine breitere Ressourcen-Basis; Umsetzung größerer Korpusstudien.

¹⁷ Um zu vermeiden, dass der Typ der ausgeblendeten Instanzen im trainierten AW systematisch falsch behandelt wird, kann die Menge der Lern-Features gezielt so angepasst werden, dass die ausgeblendeten Instanzen sich nicht von den nächstliegenden im Training gewählten Instanzen unterscheiden (im Beispiel würden die Features, die auf oberflächlichen Kasusunterscheidungen basieren, entfernt). Wenn ein erstes AW entsprechend trainiert ist, kann es wiederum mit Active-Learning- bzw. Bootstrapping-inspirierten Methoden verwendet werden, um Belege zu suchen, die nicht auf der Vereinfachung basieren.

¹⁸ Technisch ist hierfür zu prüfen, inwieweit sich die in D1 entwickelte Datenbank-Infrastruktur für die D4-interne Datenverwaltung erweitern lässt. In jedem Fall wird D4 auf das PAULA-Format aufbauen. Die Datenhaltung für den interaktiven Annotationsprozess ist separat; nur Annotationen, die endgültig manuell nachkontrolliert wurden, werden in die ANNIS-Datenbank importiert und damit auch für die Suche außerhalb der interaktiven Methodologie zur Verfügung gestellt. Dies betrifft nur einen Bruchteil der für den LAILA-Ansatz intern verwendeten Daten.

D. Interaktive Architektur, Datenhaltung und Werkzeug-Integration

2007/08: Bereitstellung der Annotationswerkzeuge; Abstimmung mit D1 zu Daten-Infrastruktur; 2009-10: Verbesserung der interaktiven Benutzbarkeit; Benutzerschnittstellen; 2011: Erweiterbarkeit; Dokumentation; Vorbereitung einer Veröffentlichung der Werkzeuge.

E. Datenannotation und Evaluation

2007/08: Spezifikation einer Evaluationsmethodologie; Annotation von Korpusdaten für systematische Vergleichsexperimente; 2009-11: Laufende Trainings- und Kontroll-Annotation im Rahmen der Korpusstudien; Evaluation auf unabhängigen Ressourcen.

F. Anwendung auf linguistische Beispielstudien zu IS

2007: Abstimmung mit Kooperationsprojekten; 2008/09: Spezifikation der relevanten Klassifikatoren zu IS-relevanten Parametern; Inspektion und Auswertung der Korpusuntersuchungsergebnisse; Untersuchung des Status von IS bei der Übersetzung und in Parallelkorpora; 2010/11: Vertiefte linguistische Studien; Erweiterung auf andere Sprachen/Phänomentypen; Überlegungen zur Modellierung von Frequenzeffekten.

3.5 Stellung innerhalb des Sonderforschungsbereichs

D4 versteht sich als Ergänzung zu den im SFB in der ersten Förderungsperiode entwickelten und umgesetzten Methoden der Datenerhebung, -aufbereitung und -annotation für ein thematisch fokussiertes typologisches Korpus zur Informationsstruktur. Die Methodologie wird z.T. in Kooperation mit anderen Teilprojekten erarbeitet; daneben bietet D4 im Sinne der Service-Funktion des D-Bereichs Unterstützung bei der Verwendung von sprachtechnologischen Verfahren auf größeren Korpora an.

Eine besonderes Ziel für D4 besteht in dem Versuch, Frequenzuntersuchungen in die linguistische IS-Forschung stärker zu integrieren als dies bisher der Fall ist. Korpusfrequenzdaten sind für die experimentellen Arbeiten aus dem C-Bereich von unmittelbarer Bedeutung (die direkte Kooperation mit C1 und C6 wurde bereits hervorgehoben). Auch für den typologischen Vergleich zum Einsatz grammatischer Mittel für die Realisierung von IS (also dem Bereich B) gewinnt in einem fortgeschrittenen Untersuchungsstadium die Einsatzhäufigkeit der unterschiedlichen Alternativen (in Abhängigkeit von diskurs-kontextuellen u.a. Faktoren) an Bedeutung. So steht beispielsweise praktisch allen Sprachen eine Cleft-Konstruktion zur Verfügung, ihre Häufigkeit zur Realisierung von Topik und Fokus schwankt jedoch stark von Sprache zu Sprache; eine detaillierte typologische Beschreibung sollte dies berücksichtigen, sofern Korpora in ausreichender Größe vorliegen. Nicht zuletzt sind in diesem Zusammenhang Korpusfrequenz-Daten für die Theoriebildung (d.h. den Bereich A) von Bedeutung. Grammatiktheorien, die graduelle Verwendungspräferenzen zwischen verschiedenen Variationsalternativen vorhersagen (beispielsweise probabilistische Versionen der Optimalitätstheorie, vgl. Boersma 1998), lassen sich auf Basis von Frequenzverteilungen in den Korpusdaten empirisch überprüfen (eine direkte Kooperation hierzu ist mit A1 zur Position von Relativsätzen im Deutschen geplant). Zentral für die aktuelle IS-Forschung ist dies insbesondere im Zusammenhang einer möglichen funktionalen Erklärung für den Einsatz grammatischer Mittel für die IS-Realisierung in Abhängigkeit von der Markiertheit (i.S.v. geringer Verwendungsfrequenz) der Alternativen.

Ein möglicher zusätzlicher Beitrag der D4-Methodologie (wie auch der in D1 laufenden statistischen Arbeiten) für die linguistische IS-Forschung kann in der Inspektion der Konfidenz/Unsicherheit liegen, mit der die auf Beispielinstanzen trainierten automatischen Verfahren bestimmte Daten klassifizieren: dies kann zumindest Anhaltspunkte für besonders geeignete und eher problematische theoretische Kategorien liefern.

Nicht zuletzt bereitet der Einsatz diverser sprachtechnologischer Werkzeuge und Techniken für Aspekte der IS-Analyse mögliche Anwendungen von Ergebnissen aus dem SFB in einer geplanten dritten Förderungsphase vor.

3.6 Abgrenzung gegenüber anderen geförderten Projekten des/der Teilprojektleiter/ Teilprojektleiterinnen

Die für D4 geplanten Forschungsaufgaben sind in mehrererlei Hinsicht komplementär zum laufenden PTOLEMAIOS-Projekt. Während letzteres im Kern auf grundlagenorientierte Untersuchungen zur Lernbarkeit von Grammatiken fokussiert und neben der implizit in Parallelkorpora enthaltenen Strukturinformation keine linguistischen Ressourcen in den Lernprozess einbezieht, sollen in D4 existierende Werkzeuge möglichst umfassend ausgenutzt werden, um bei der halbautomatischen Korpusanalyse ein Ergebnis zu erzielen, das einerseits den manuellen Annotierungs- bzw. Kontrollaufwand sorgfältig kanalisiert und andererseits die hohen Qualitätsanforderungen einer linguistischen Verwertbarkeit erfüllt. D4 zielt auf die Analyse von IS und IS-relevanten Kontext-Parametern ab, während der Gegenstand von PTOLEMAIOS die Lernbarkeit von kernsyntaktischen Regeln und Prinzipien ist. Parallelkorpora spielen in D4 die Rolle eines Typs unter verschiedenen nutzbaren Ressourcen, während PTOLEMAIOS sowohl für die Grammatikinduktion als auch für den zweiten Schwerpunkt maschinelle Übersetzung sich ausschließlich auf Parallelkorpora konzentriert. Die interaktive Methode der Korpusnutzung in D4 schließlich steht ebenfalls in Kontrast zu den unüberwachten Verfahren in PTOLEMAIOS.

Trotz der deutlich unterschiedlichen Zielrichtung kann durch bestimmte methodologische Überschneidungen mit Synergie-Effekten gerechnet werden, insbesondere für technische Aspekte die Verwendung von Parallelkorpora in D4, und wie erwähnt zu den eingesetzten ML-Techniken (Feature-Selektion, hierarchische Abhängigkeiten). Dies ist für die implementierungsintensiven computerlinguistischen Projekte eine wichtige Erleichterung, ohne die der Skopus für D4 reduziert werden müsste.

Literatur

- Ahrens, Barbara. 2003. Prosodie beim Simultandolmetschen. Peter Lang.
- Baroni, Marco and Evert, Stefan. To appear. Statistical methods for corpus exploitation. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, chapter 38. Mouton de Gruyter, Berlin.
- Baldrige, Jason, and Miles Osborne. 2003. Active learning for HPSG parse selection. In *Proceedings of the 7th Conference on Natural Language Learning*.
- Bender, E. M., D. Flickinger and S. Oepen. 2002. The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. Carroll, J., N. Oostdijk, and R. Sutcliffe, eds. *Proc. of the Workshop on Grammar Engineering and Evaluation at the 19th Int'l Conf. on Comp. Ling.*, Taipei, Taiwan. 8–14.
- Boersma, Paul. 1998. *Functional Phonology. Formalizing the interactions between articulatory and perceptual drives*. PhD thesis, University of Amsterdam.
- Bader, Markus and Jana Häussler. 2006. Weighting the Constraints on Word-order. Variation in German. Slides for presentation at QITL 2 (Quantitative Investigations in Theoretical Linguistics 2), Osnabrück 2006. <http://www.cogsci.uni-osnabrueck.de/~qitl/>
- Becker, M., A. Bredenkamp, B. Cysmann, J. Klein. 2003. Annotation of Error Types for a German Newsgroup Corpus. In A. Abeille (ed): *Treebanks. Building and Using Parsed Corpora*, Kluwer Academic Publishers, The Netherlands.
- Becker, M., B. Hachey, B. Alex and C. Grover. 2005. Optimising Selective Sampling for Bootstrapping Named Entity Recognition. In *Proceedings of the ICML-2005 Workshop on Learning with Multiple Views*, Bonn, Germany.
- Becker, M., M. Osborne. 2005. A Two-Stage Method for Active Learning of Statistical Grammars. In *Proceedings of IJCAI 2005*, Edinburgh, UK.
- Becker, M. and E. Pecourt. 2002. Anaphora Resolution Using a Topological Parser. In *Proceedings of DAARC 2002*, Lisbon, Portugal.
- Becker, M. and A. Frank. 2002. A Stochastic Topological Parser of German. In *Proceedings of COLING 2002*, Taipei, Taiwan.
- Blum, Avrim, and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory*. Morgan Kaufmann Publishers.
- Brants, Thorsten. 1999. *Tagging and Parsing with Cascaded Markov Models - Automation of Corpus Annotation*. Saarbrücken Dissertations in Computational Linguistics and Language Technology, Volume 6. DFKI and Saarland University, Saarbrücken, Germany.

- Brants, Thorsten. 2000. TnT - A Statistical Part-of-Speech Tagger. In Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, Seattle, WA.
- Bresnan, Joan, Shipra Dingare, and Christopher Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In M. Butt and T. H. King (eds.), Proceedings of the LFG 01 Conference. CSLI Publications.
- Bresnan, Joan, Ashwini Deo, and Devyani Sharma. 2006. Typology in Variation: A Probabilistic Approach to be and n't in the Survey of English Dialects. Ms. Erscheint in English Language and Linguistics.
- Bresnan Joan, and Jennifer Hay. 2006. Gradient Grammar: An Effect of Animacy on the Syntax of *give* in Varieties of English. Ms. Stanford University.
- Brill, Eric. 1993. A Corpus-Based Approach to Language Learning. PhD Dissertation, University of Pennsylvania.
- Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar Project. In Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation. 1–7.
- Butt, Miriam and Tracy Holloway King. 2002. Urdu and the Parallel Grammar Project. In Proceedings of COLING-2002 Workshop on Asian Language Resources and International Standardization. pp. 39-45.
- Cohn, David A., Zoubin Ghahramani, and Michael I. Jordan. 1995. Active learning with statistical models. In G. Tesauero, D. Touretzky, and T. Leen, editors, Advances in Neural Information Processing Systems, volume 7, pages 705–712. The MIT Press.
- Charniak, Eugene. 2000. A maximum entropy-inspired parser. In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, pp. 132–139, Seattle, Washington.
- Collins, Michael. 1997. Three generative, lexicalised models for statistical parsing. In Proceedings of the 35th Annual Meeting of the ACL. 16–23.
- Daelemans, Walter, Jakub Zavrel, Peter Berck and Steven Gillis. 1996. MBT: A Memory-Based Part of Speech Tagger-Generator.. in: E. Ejerhed and I. Dagan (eds.) Proceedings of the Fourth Workshop on Very Large Corpora, Copenhagen, Denmark, 14-27.
- De Sutter, G., D. Speelman and D. Geeraerts. Erscheint. Detecting and balancing determinants of word order variation in Dutch clause final verb clusters. To appear in Linguistics.
- Dubey, Amit. 2004. Statistical Parsing for German: Modeling syntactic properties and annotation differences. PhD thesis, Saarland University, Germany.
- Eckert, Miriam & Michael Strube. 2000. Dialogue acts, synchronizing units and anaphora resolution. Journal of Semantics, 17(1):51–89.
- Eckle Judith, and Ulrich Heid. 1996. Extracting raw material for a German subcategorization lexicon from newspaper text. In: Proceedings of the 4th International Conference on Computational Lexicography, COMPLEX'96. Budapest, Hungary.
- Erk, Katrin and Sebastian Pado. 2006. Shalmaneser - a flexible toolbox for semantic role assignment. In Proceedings of LREC-06, Genoa.
- Erk, Katrin, Andrea Kowalski, Sebastian Pado and Manfred Pinkal. 2003. Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation. Proceedings of ACL 2003, Sapporo.
- Evert, Stefan. 2004. The statistical analysis of morphosyntactic distributions. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), pages 1539 - 1542, Lisbon, Portugal.
- Evert, Stefan and Krenn, Brigitte (2005). Using small random samples for the manual evaluation of statistical association measures. Computer Speech & Language 19(4), 450 - 466.
- Frank, A., M. Becker, B. Crysmann, B. Kiefer, and U. Schäfer, 2003. Integrated Shallow and Deep Parsing: TopP meets HPSG. In Proceedings of ACL 2003, Sapporo, Japan.
- Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. In Jennifer Spender, Anders Eriksson, and Östen Dahl (Eds.), Proceedings of the Stockholm Workshop on 'Variation within Optimality Theory'. April 26-27, 2003 at Stockholm Univ. Sweden, pp. 111–120.
- Hachey, B., B. Alex, M. Becker, 2005. Investigating the Effects of Selective Sampling on the Annotation Task. In Proceedings of CoNLL 2005, Ann Arbor, USA.
- Hamp, Birgit and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In: Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications". Madrid, 1997.
- Henderson, John C. and Brill, Eric. 1999. Exploiting Diversity in Natural Language Processing: Combining Parsers. In Proceedings of the Fourth Conference on Empirical Methods in Natural Language Processing (EMNLP-99), pages 187-194.
- Hwa, Rebecca, Philip Resnik, and Amy Weinberg. 2002. Breaking the resource bottleneck for multilingual parsing. In Proceedings of LREC.
- Karttunen, Lauri, J-P. Chanod, G.Grefenstette, and A.Schiller. 1996. Regular expressions for language engineering. Natural Language Engineering 2(4), pp. 305-328.

- Jäger, Gerhard. 2006. Maximum Entropy Models and Stochastic Optimality Theory, to appear in Jane Grimshaw, Joan Maling, Chris Manning, Jane Simpson, and Annie Zaenen (eds.), *Architectures, Rules, and Preferences: A Festschrift for Joan Bresnan*, CSLI Publications, Stanford.
- Kehler, Andrew. 1997. Probabilistic coreference in information extraction. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, 163–173.
- Kermes, Hannah and Evert, Stefan (2001). Exploiting large corpora: A circular process of partial syntactic analysis, corpus query and extraction of lexicographic information. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, and S. Khoja (eds.), *Proceedings of the Corpus Linguistics 2001 Conference*, pages 332 - 340, Lancaster. UCREL.
- Kermes, Hannah and Evert, Stefan. 2002. YAC - a recursive chunker for unrestricted German text. In M. G. Rodriguez and C. P. Araujo (eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 1805–1812.
- Kermes, H. and Heid, U. 2003. Using chunked corpora for the acquisition of collocations and idiomatic expressions. In *Proceedings of COMPLEX 2003*.
- Klein, Dan and Christopher D. Manning. 2002. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, December 2002.
- Koehn, Philipp. 2002. *Europarl: A multilingual corpus for evaluation of machine translation*. Ms., University of Southern California.
- Lin, Dekang. 1998. Dependency-based evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems*, Granada, Spain.
- Nivre, J., Hall, J., Nilsson, J., Eryigit, G. and Marinov, S. (2006) Labeled Pseudo-Projective Dependency Parsing with Support Vector Machines. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL)*.
- Müller, C., S. Rapp, and M. Strube. Applying co-training to reference resolution. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 352–359, Morristown, NJ, USA, 2001. Association for Computational Linguistics.
- Ng, V. and C. Cardie. Improving machine learning approaches to coreference resolution. 2001. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 104–111.
- Ng, V. and C. Cardie. Bootstrapping coreference classifiers with multiple machine learning algorithms. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 113–120.
- Pado, Sebastian and Mirella Lapata. 2005. Cross-linguistic projection of role-semantic information. *Proceedings of HLT/EMNLP 2005*, Vancouver, Canada.
- Pradhan, Sameer, Wayne Ward, Kadri Hacioglu, James Martin, and Dan Jurafsky. 2005. Semantic role labeling using different syntactic views. In *Proceedings of the Association for Computational Linguistics 43rd annual meeting (ACL-2005)*, Ann Arbor, MI.
- Ramshaw, L. and M. Marcus. 1995. Text Chunking using Transformation-Based Learning. *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94.
- Ellen Riloff and Jessica Shephard. 1997. A corpus-based approach for building semantic lexicons. In Claire Cardie and Ralph Weischedel, eds., *Proceedings of EMNLP*, 117–124.
- Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*.
- Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.
- Seung, H. S., Manfred Opper, and Haim Sompolinsky. 1992. Query by committee. In *Computational Learning Theory*, 287–294.
- Soon, Wee Meng, Hwee Tou Ng and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.
- Samuelson, C. and Voutilainen, A. 1997. Comparing a Linguistic and a Stochastic Tagger. *Proceedings of joint ACL/EACL 1997*. Madrid, Spain.
- Steedman, M., M. Osborne, A. Sarkar, S. Clark, R. Hwa, J. Hockenmaier, P. Ruhlen, S. Baker, and J. Crim. 2003. Bootstrapping statistical parsers from small datasets. In *Proceedings of EACL 2003*, Budapest, Hungary.
- Thompson, C. A., Califf, M. E., Mooney, R. J. 1999. Active learning for natural language parsing and information extraction In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99)*, 406-414.
- Ule T. and Müller F. H. 2001. KaRoPars: Ein System zur linguistischen Annotation großer Textkorpora des Deutschen. In *Proceedings of the Workshop Werkzeuge zur automatischen Analyse und Verarbeitung von Texten: Formate, Tools, Software-Systeme*, Trier, Germany.
- Wasow, Thomas, T. Florian Jäger, and David Orr. *Erscheint. Lexical Variation in Relativizer Frequency*. *Erscheint in Horst Simon and Heike Wiese (eds.): Proceedings of the workshop on exceptions in grammar, DGfS annual meeting 2005*.
- Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189–196, Cambridge, MA.

Yarowsky, D., G. Ngai, and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In Proceedings of HLT 2001, First International Conference on Human Language Technology Research, pp. 161–168.